Hybrid enrichment –
Anchored phylogenomics

Collection of High-Throughput Data

- Targeted Amplicon Sequencing or Parallel Tagged Sequencing
- Multiplex PCR
- Massively Parallel Uniplex PCR
- Reduced Representation Library or RAD Sequencing
- Transcriptome Sequencing
- **Hybrid Enrichment**

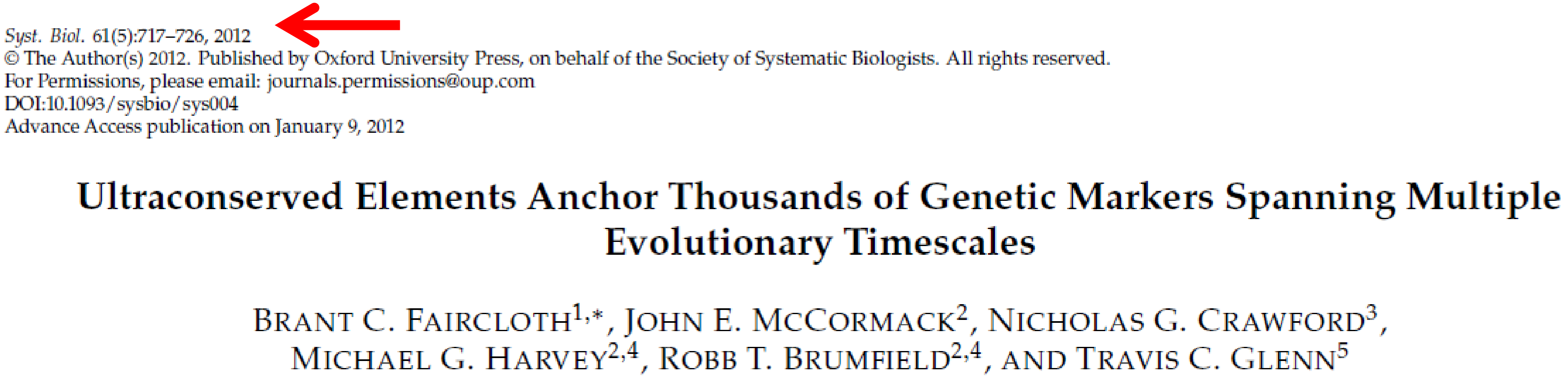
- **BUT: more data = better data ??**

Table 1
Methods of sample preparation for using NGS in phylogeography and phylogenetics.

Method	Other names or variants	Literature method	Literature examples	Benefits	Drawbacks	Best application
Amplicon sequencing	Multiplex PCR, parallel tagged sequencing	Binladen et al. (2007), Meyer et al. (2008), Tewhey et al. (2009b)	Chan et al. (2010), Griffin et al. (2011), Gunnarsdóttir et al. (2011); Morin et al. (2010), Parks et al. (2009)	Highly targeted. Results in nearly complete data matrices. Needed coverage easy to calculate. Circumvents individual sequencing reactions and phasing nuclear loci compared to Sanger sequencing	Requires PCR of each individual at each locus	Small- to medium-scale projects targeting a limited number of genes
Restriction-digest	Double-digest genome reduction, RAD sequencing (RAD-seq), complexity reduction of multilocus sequences (CRoPS), Genotyping by Sequencing (GBS)	Baird et al. (2008), Davey et al. (2011)	Andolfatto et al. (2011), Amaral et al. (2009), Bers et al. (2010), Emerson et al. (2010), Gompert et al. (2010), Hohenlohe et al. (2011), Hyten et al. (2010a,b), Kerstens et al. (2009), Ramos et al. (2009); Sánchez et al. (2009); Van Orsouw et al. (2007); Van Tassell et al. (2008), Wiedmann et al. (2008); Williams et al. (2010)	Broad, random genomic sampling of thousands of independent genomic regions. Requires no prior genomic resources whatsoever	Reproducibility and throughput may be limited by gel extraction step. Not targeted, thus coverage can be difficult to estimate. Null alleles could skew diversity estimates	Intraspecific studies of recent divergence
Target enrichment	Sequence capture, targeted resequencing, primer extension capture (PEC)	Albert et al. (2007), Gnirke et al. (2009), Hodges et al. (2007), Okou et al. (2007), Tewhey et al. (2009a), Maricic et al. (2010)	Briggs et al. (2009; Faircloth et al. in press)	Rapid collection of thousands of loci without individual PCR	Requires some prior genomic knowledge, but not necessarily a sequenced genome	Phylogenomics at taxonomic levels above at and above the species level
Transcriptome	RNA-seq	Morin et al. (2008); Marioni et al. (2008)	Barbazuk and Schnable (2011), Cánovas et al. (2010), Chepelev et al. (2009); Geraldes et al. (2011), Hittinger et al. (2010)	Can leverage data from expression studies	Skewed read distributions can outstrip coverage, making it difficult to find orthologous loci	Leveraging existing cDNA libraries

Comparison of Methods

	Broad Taxonomic Application	Efficient With Non-Model Species	Useable across Tree of Life	Long Loci	Minimal Missing Data	Broad Phylogenetic Informativeness	Utilizes DNA	Accommodates Degraded Samples	Efficient With Large Genome	Short Development Time	Short Data Collection Time	Short Data Assembly Time	Robust to Substitutions	Cost Effective
Hybrid Enrichment	●	●	●	●	◒	●	●	●	●	◒	●	◒	◒	●
Whole Genome	●	●	●	●	●	●	●	●	🔴	●	●	🔴	●	🔴
Transcriptome	●	●	●	●	◒	◒	🔴	🔴	●	●	●	◒	◒	🔴
RRL & RAD	🔴	●	●	🔴	🔴	🔴	●	◒	◒	◒	●	●	🔴	●
Multiplex PCR	◒	◒	●	●	◒	🔴	●	◒	●	🔴	◒	●	🔴	◒
Targeted Amplicon	◒	◒	●	●	●	🔴	●	◒	●	🔴	🔴	●	🔴	◒
Massively Uniplex PCR	◒	◒	●	●	●	🔴	●	◒	●	🔴	●	●	🔴	◒

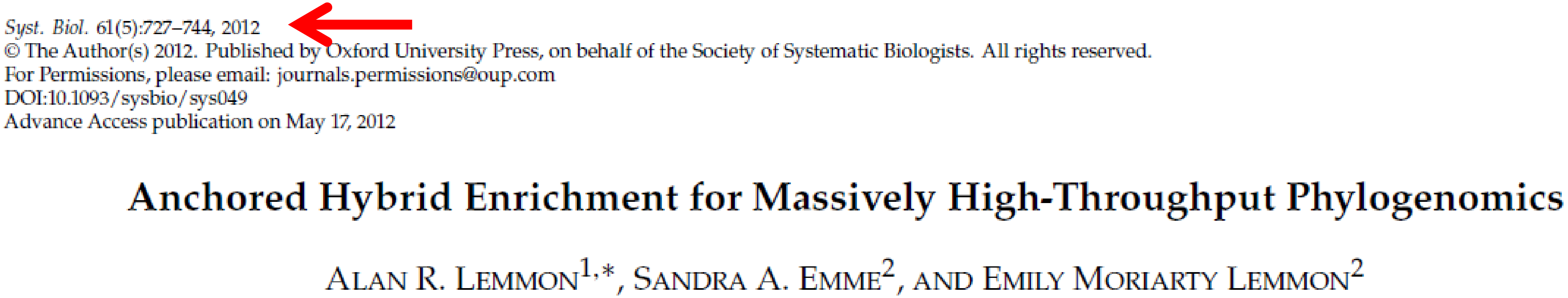


Syst. Biol. 61(5):717–726, 2012

DOI:10.1093/sysbio/sys004
Advance Access publication on January 9, 2012

Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales

Brant C. Faircloth[1,*], John E. McCormack[2], Nicholas G. Crawford[3], Michael G. Harvey[2,4], Robb T. Brumfield[2,4], and Travis C. Glenn[5]



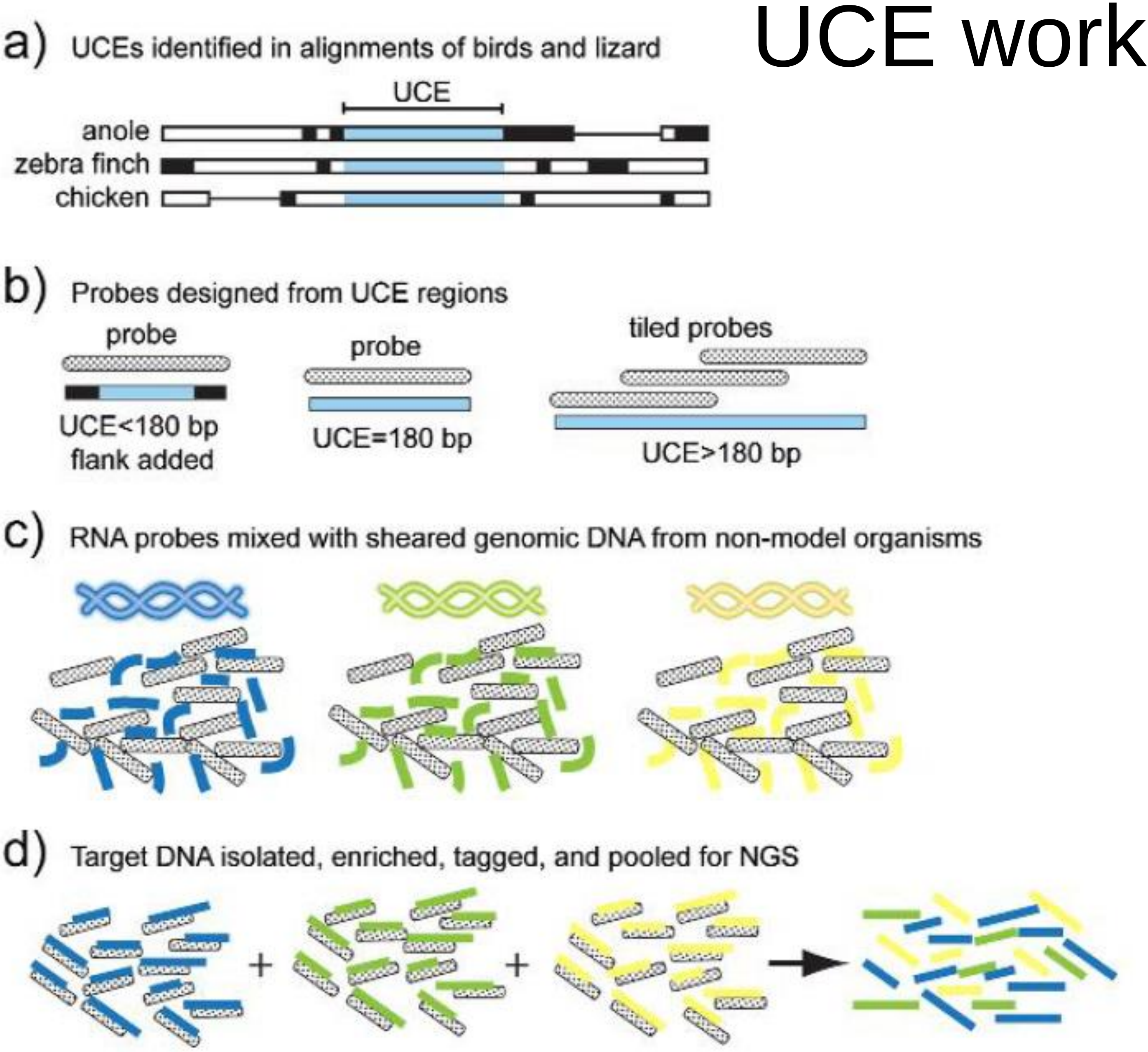
Syst. Biol. 61(5):727–744, 2012

DOI:10.1093/sysbio/sys049
Advance Access publication on May 17, 2012

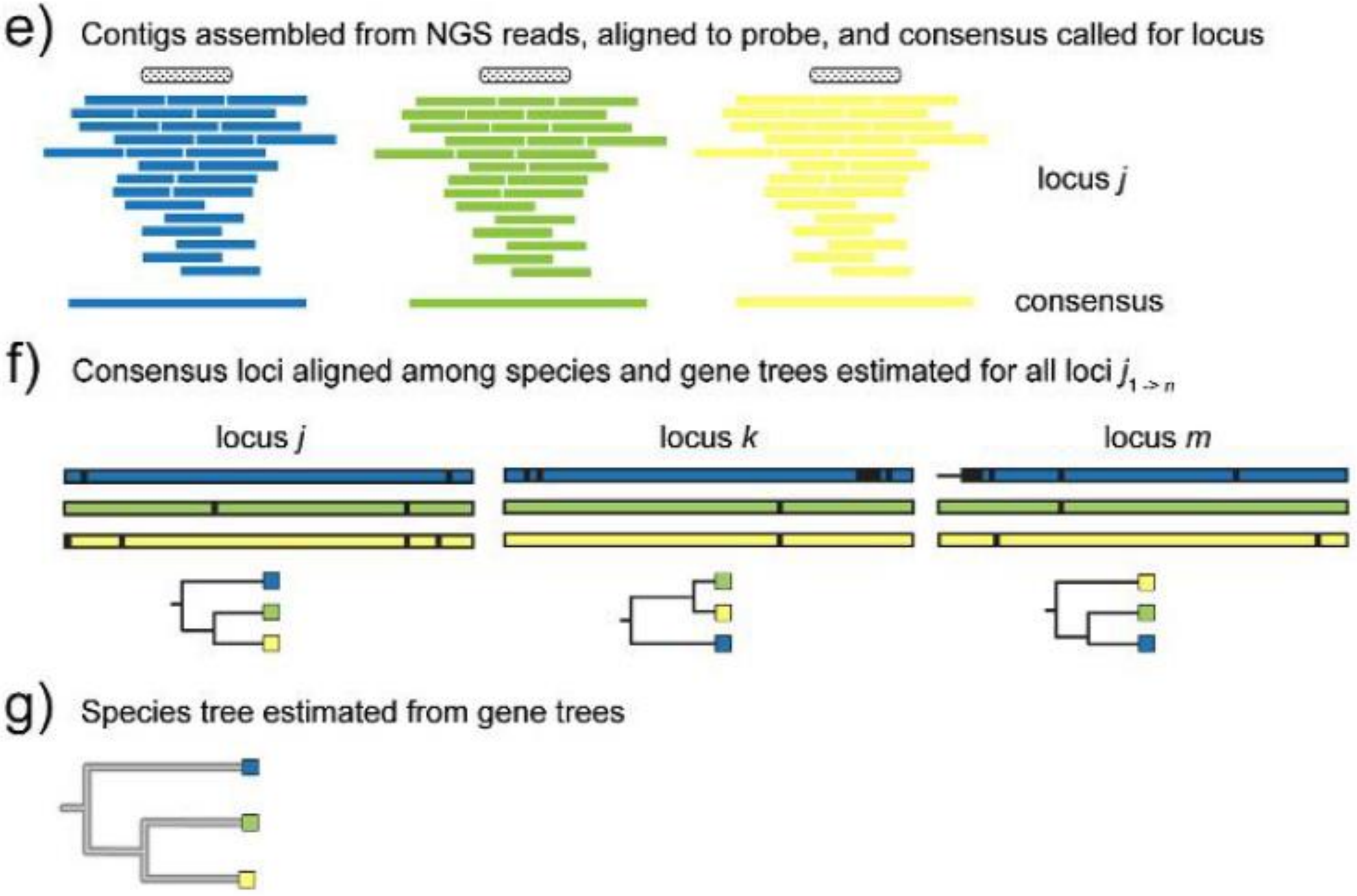
Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics

Alan R. Lemmon[1,*], Sandra A. Emme[2], and Emily Moriarty Lemmon[2]

UCE workflow


a) UCEs identified in alignments of birds and lizard
UCE
anole
zebra finch
chicken
b) Probes designed from UCE regions
probe
UCE<180 bp
flank added
probe
UCE=180 bp
tiled probes
UCE>180 bp
c) RNA probes mixed with sheared genomic DNA from non-model organisms
d) Target DNA isolated, enriched, tagged, and pooled for NGS
+
+

UCE workflow

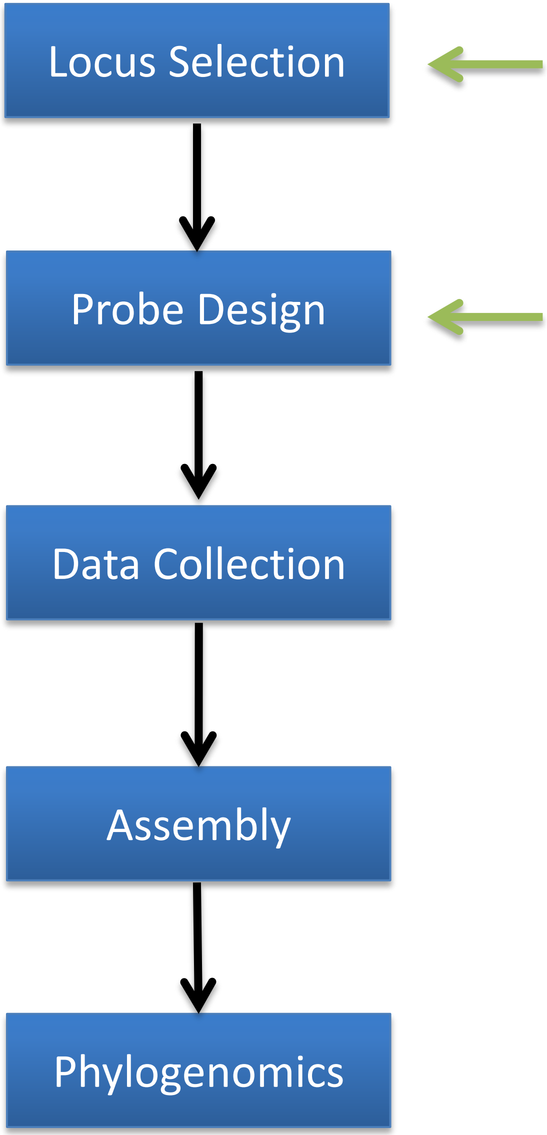


e) Contigs assembled from NGS reads, aligned to probe, and consensus called for locus

f) Consensus loci aligned among species and gene trees estimated for all loci $j_{1 \to n}$

g) Species tree estimated from gene trees

Anchored Phylogenomics Workflow



Locus Selection

Probe Design

Data Collection

Assembly

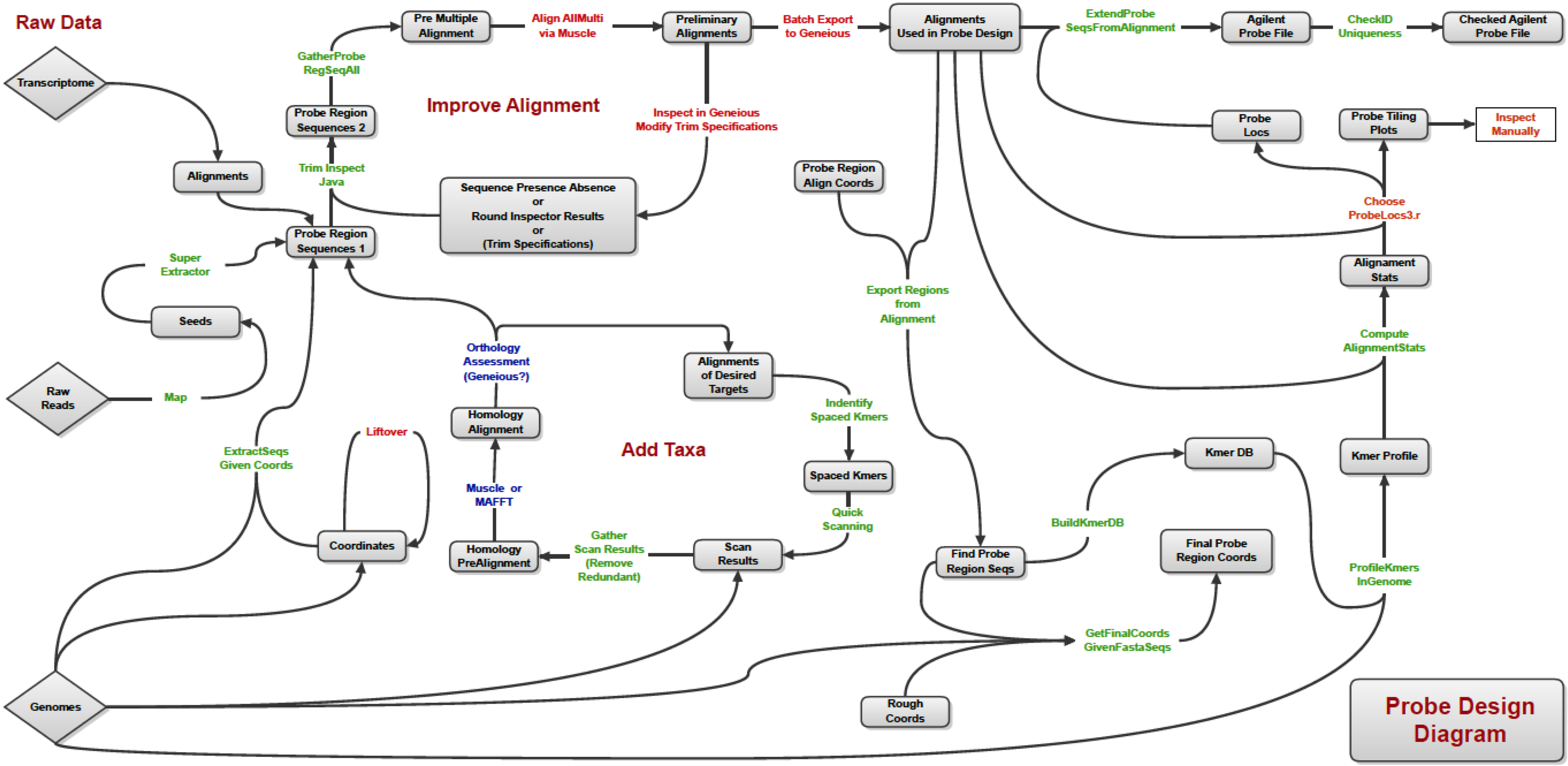
Phylogenomics

Locus Selection & Probe Design

- Locus Selection
  - ~500 “single copy” loci (typically long exons)
  - Conserved element (~20% divergence required)
  - Adjacent to less conserved regions
  - Loci are selected based on broad taxonomic group (e.g., vertebrates)

- Probe Design
  - Incorporate sufficient number of lineages
  - Tile probes across conserved region
  - Goal is to capture ~1500bp regions
  - Probe sets are designed for project-specific clade

Probe Design Workflow



Raw Data

Transcriptome

Raw Reads

Genomes

Alignments

Seeds

Super Extractor

Map

ExtractSeqs Given Coords

Coordinates

Liftover

Probe Region Sequences 1

Trim Inspect Java

Probe Region Sequences 2

GatherProbe RegSeqAll

Pre Multiple Alignment

Align AllMulti via Muscle

Preliminary Alignments

Improve Alignment

Inspect in Geneious Modify Trim Specifications

Sequence Presence Absence or Round Inspector Results or (Trim Specifications)

Batch Export to Geneious

Alignments Used in Probe Design

ExtendProbe SeqsFromAlignment

Agilent Probe File

CheckID Uniqueness

Checked Agilent Probe File

Probe Locs

Probe Tiling Plots

Inspect Manually

Choose ProbeLocs3.r

Alignament Stats

Compute AlignmentStats

Kmer Profile

ProfileKmers InGenome

Probe Region Align Coords

Export Regions from Alignment

Orthology Assessment (Geneious?)

Alignments of Desired Targets

Indentify Spaced Kmers

Spaced Kmers

Quick Scanning

Scan Results

Gather Scan Results (Remove Redundant)

Homology PreAlignment

Muscle or MAFFT

Homology Alignment

Add Taxa

Find Probe Region Seqs

BuildKmerDB

Kmer DB

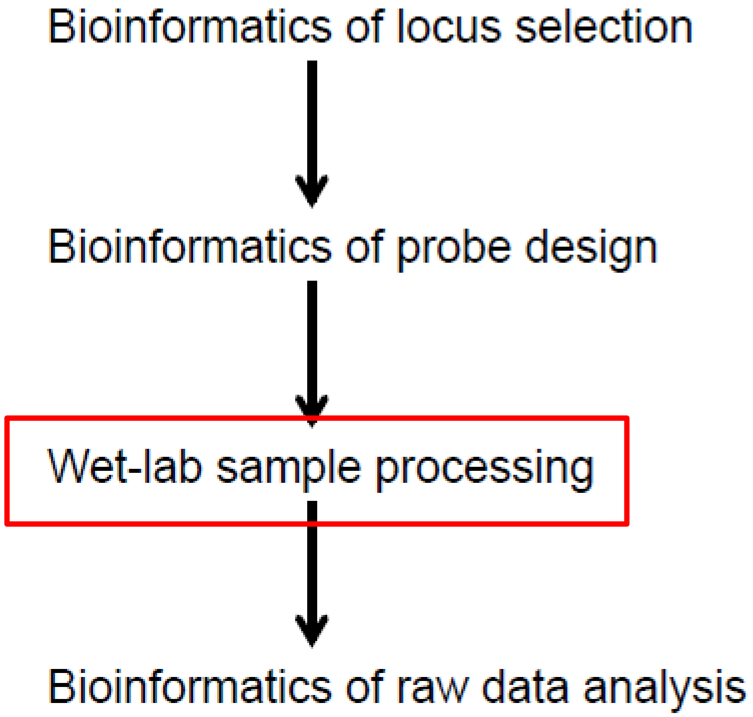
Final Probe Region Coords

GetFinalCoords GivenFastaSeqs

Rough Coords

Probe Design Diagram

Phases of Hybrid Enrichment for Phylogenomics



Bioinformatics of locus selection

Choose target loci using genomes

Bioinformatics of probe design

Tile probes across target regions

Wet-lab sample processing

Genomic DNA→ raw sequence reads

Bioinformatics of raw data analysis

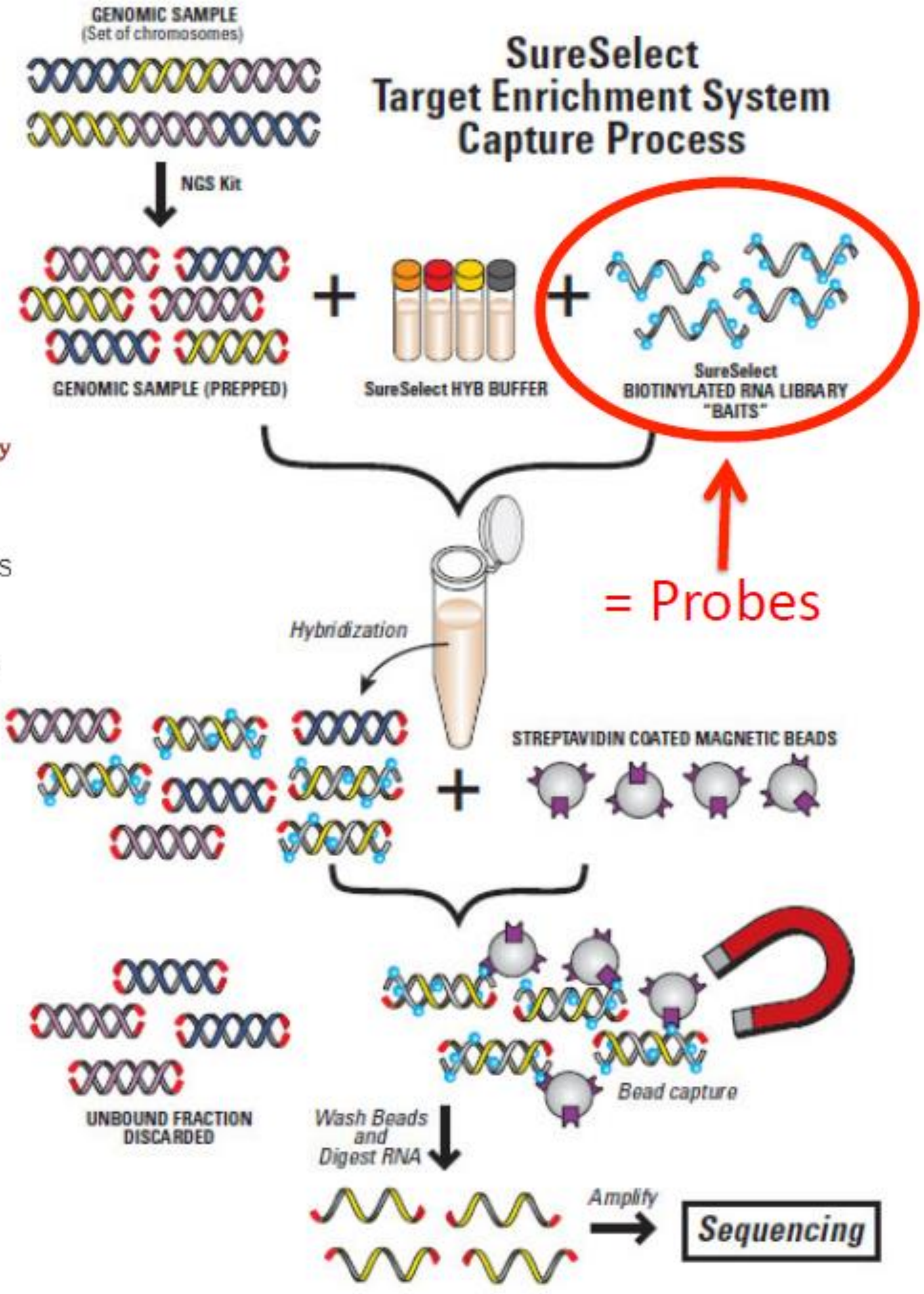
Raw sequence reads→
phased alleles, alignments, models of evolution, concordance analyses, and phylogenies

Solution-Phase Hybridization

nature biotechnology

Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing

Andreas Gnirke[1], Alexandre Melnikov[1], Jared Maguire[1], Peter Rogov[1], Emily M LeProust[2], William Brockman[1,5], Timothy Fennell[1], Georgia Giannoukos[1], Sheila Fisher[1], Carsten Russ[1], Stacey Gabriel[1], David B Jaffe[1], Eric S Lander[1,3,4] & Chad Nusbaum[1]



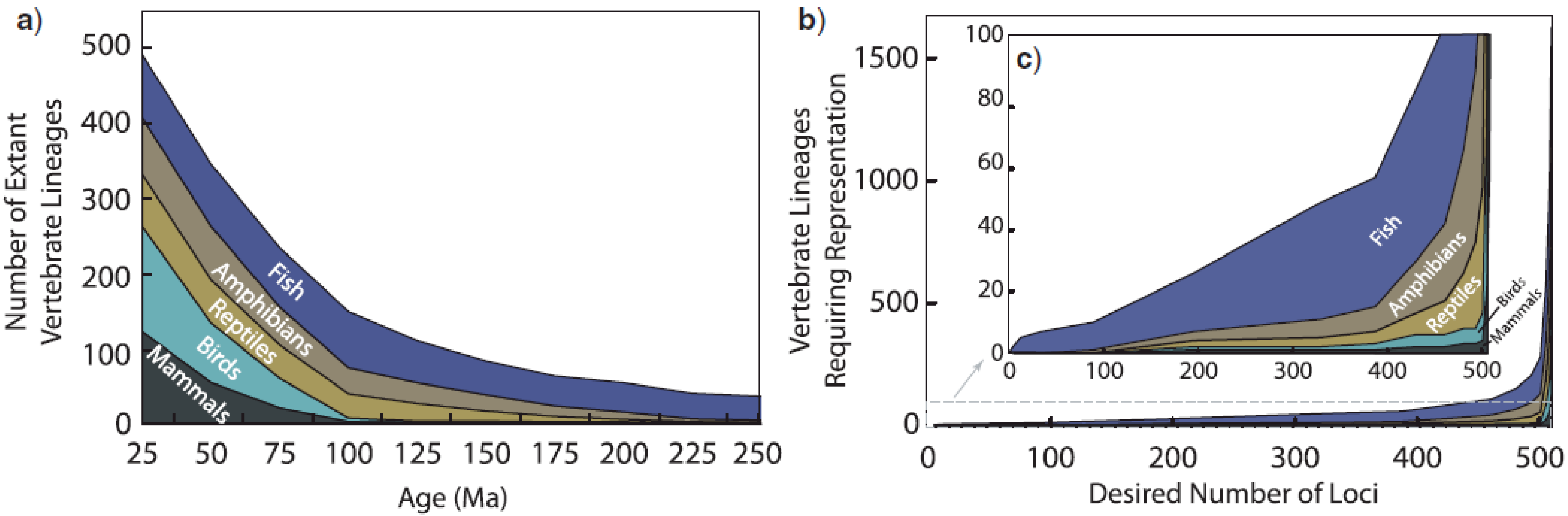
Gnirke et al. 2009

Hybrid Enrichment

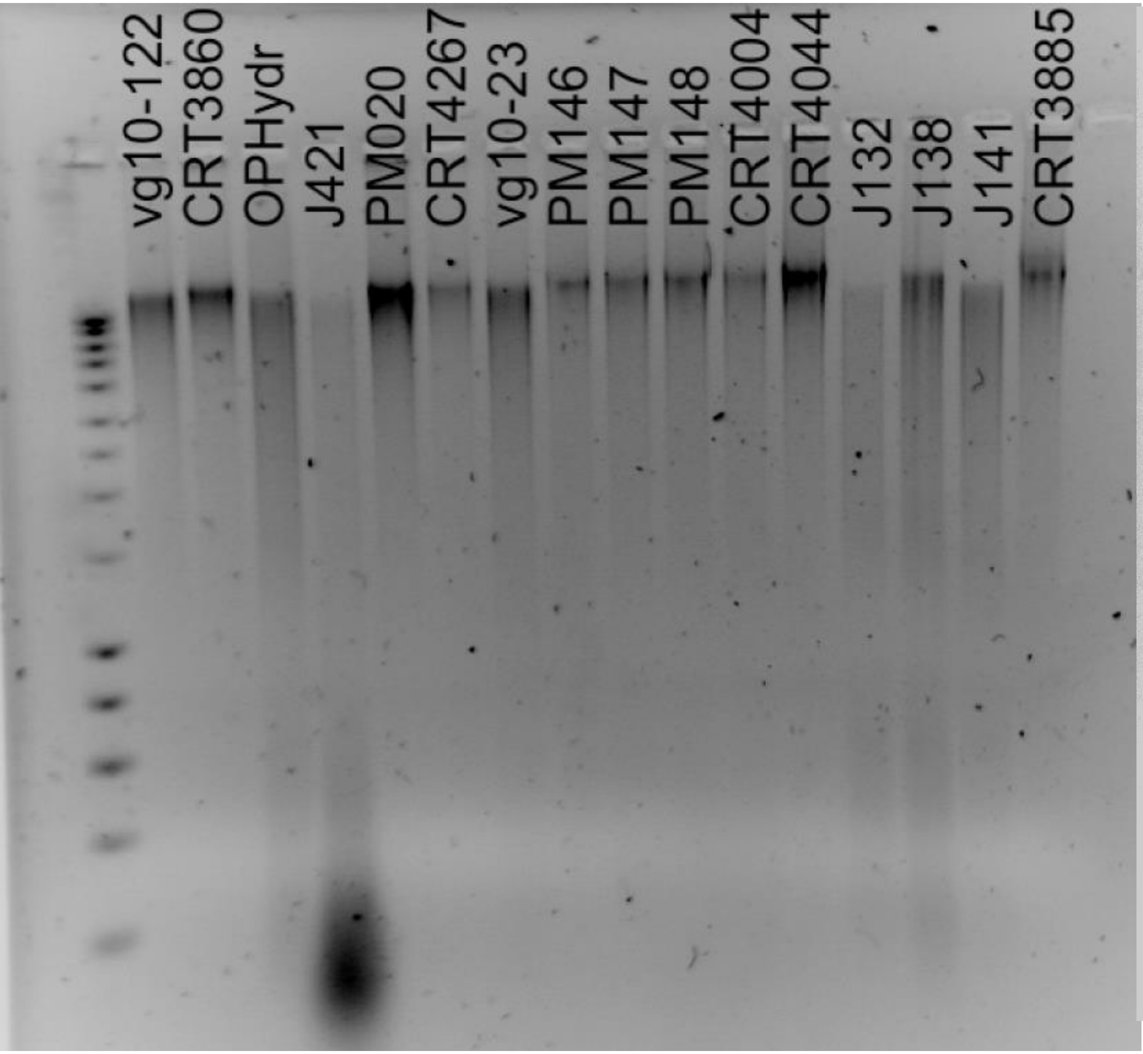
- Disadvantages:
  - Large equipment investment for small operations
    - Equipment required (Bioanalyzer, Covaris or other sonicator, qPCR machine, etc.)
  - Bioinformatic training required for locus selection, probe design, analysis of raw data
  - Substantial investment in reagents
    - Indexes for library preparation
    - Other library preparation reagents
    - Hybrid enrichment kit

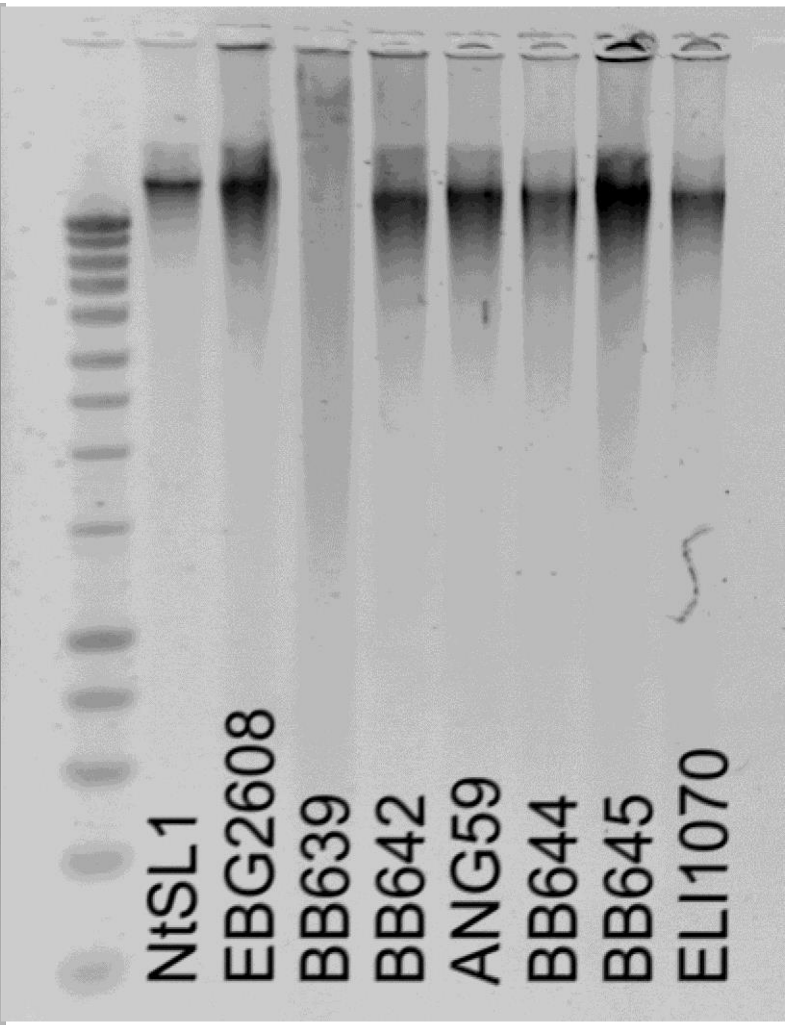
Hybrid Enrichment

- Advantages:
  - Large quantity of data
  - Complete data matrices! Can trim out all missing data and still have more than enough
  - Fast data collection (DNA to phylogeny 2 weeks)
  - Works on degraded samples and ancient DNA
  - DNA starting material (RNA not needed)
  - Can use single probe design (kit) for broad taxonomic group (e.g., Vertebrata)
  - No problem for non-model systems


a)
Number of Extant Vertebrate Lineages
500
400
300
200
100
0
Fish
Amphibians
Reptiles
Birds
Mammals
25
50
75
100
125
150
175
200
225
250
Age (Ma)
b)
Vertebrate Lineages Requiring Representation
1500
1000
500
0
0
100
200
300
400
500
Desired Number of Loci
c)
100
80
60
40
20
0
Fish
Amphibians
Reptiles
Birds
Mammals
0
100
200
300
400
500

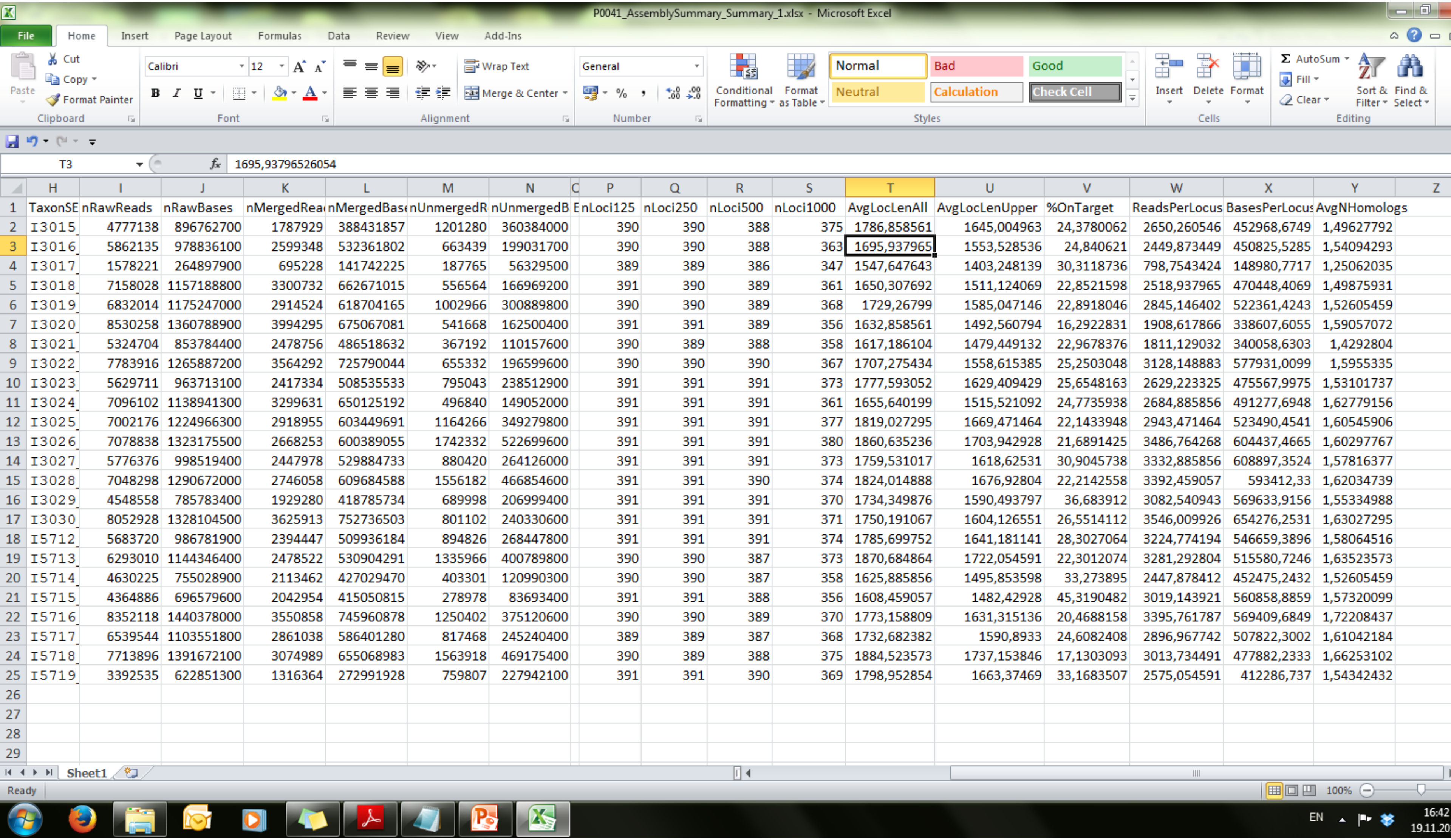
Own experiences: samples → trees





Requirement: > 1.5- 2.0 micrograms of RNA-free DNA, quantified by Qubit

Output summary

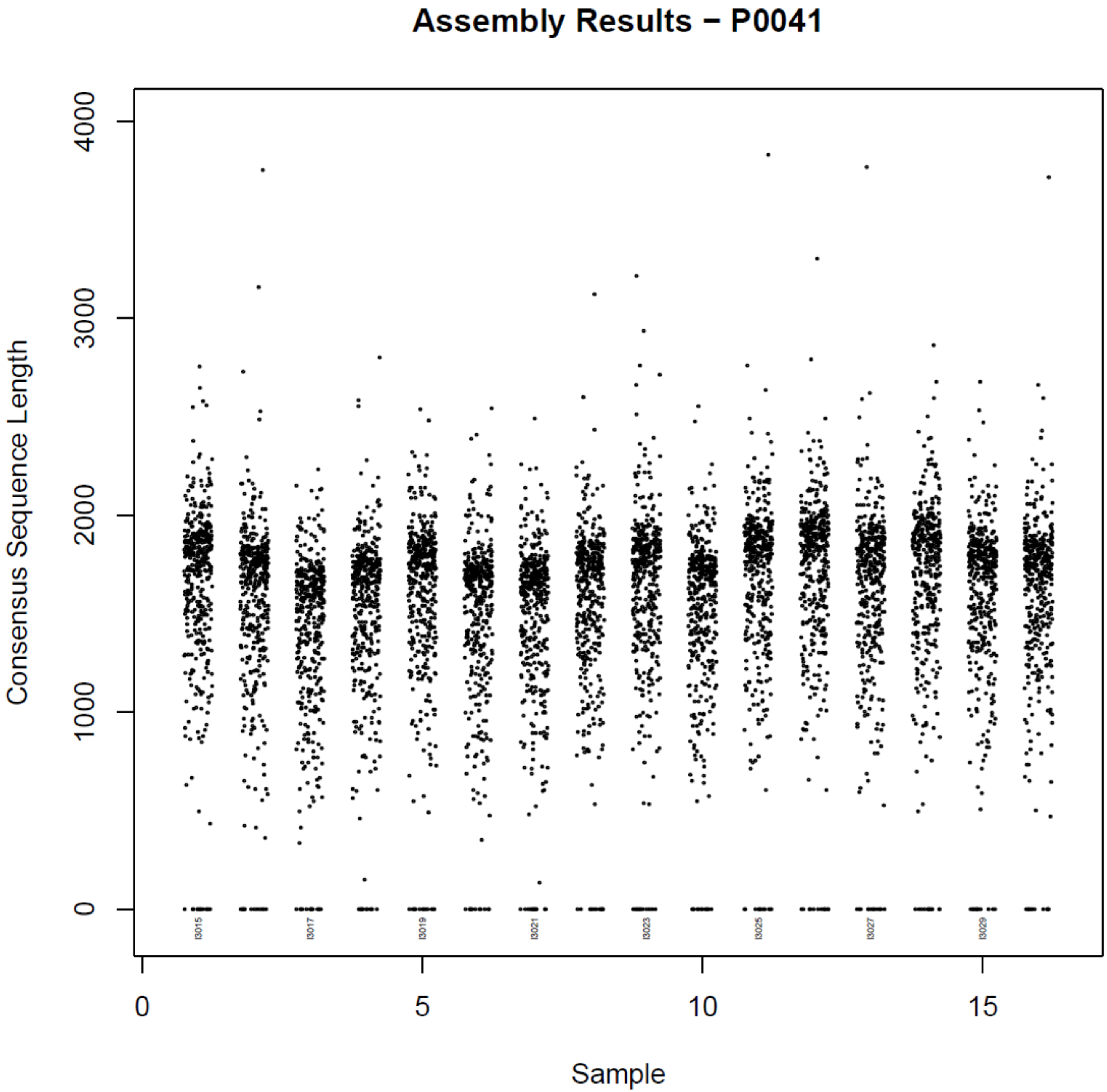


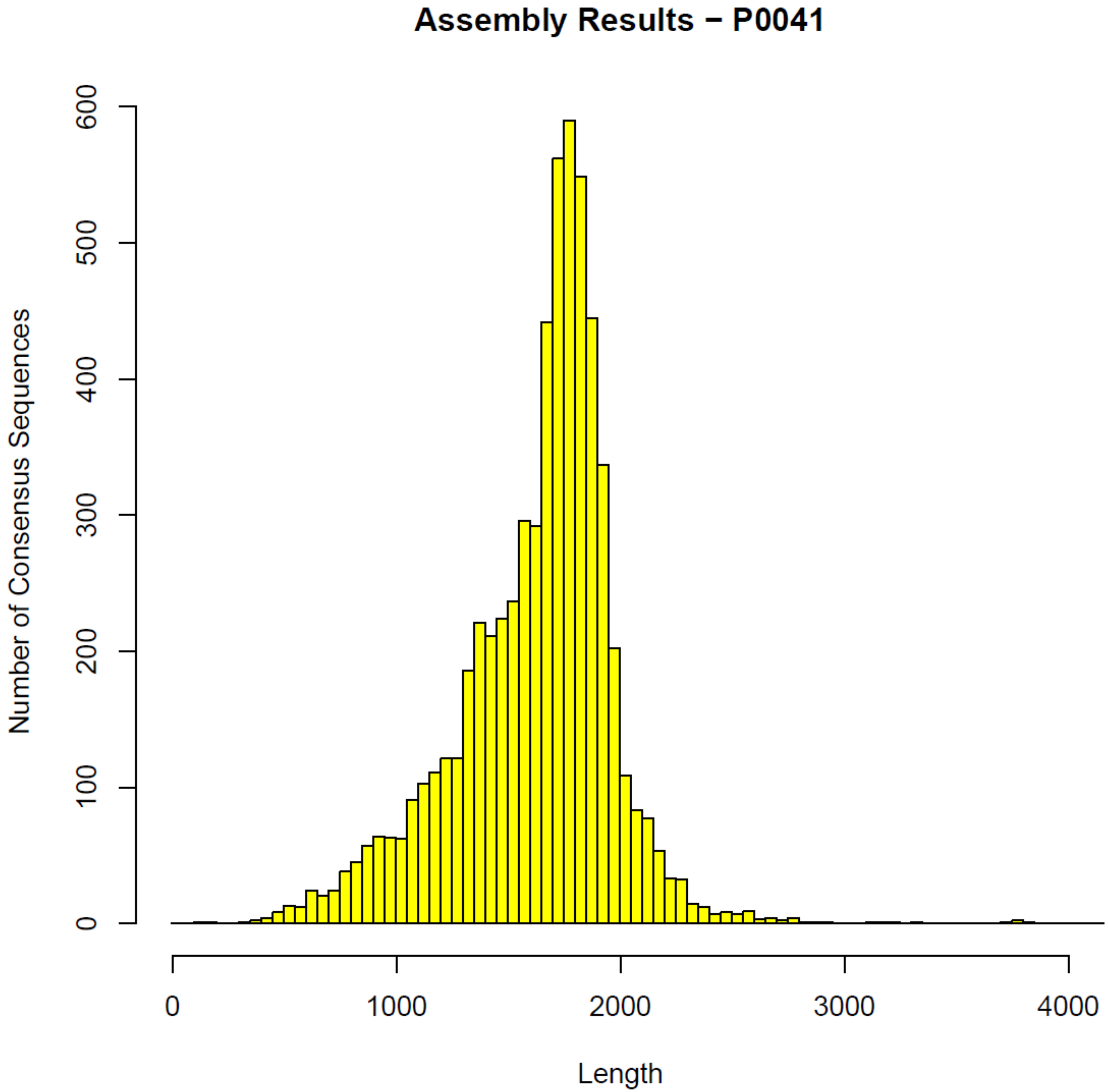
T3 fx 1695,93796526054

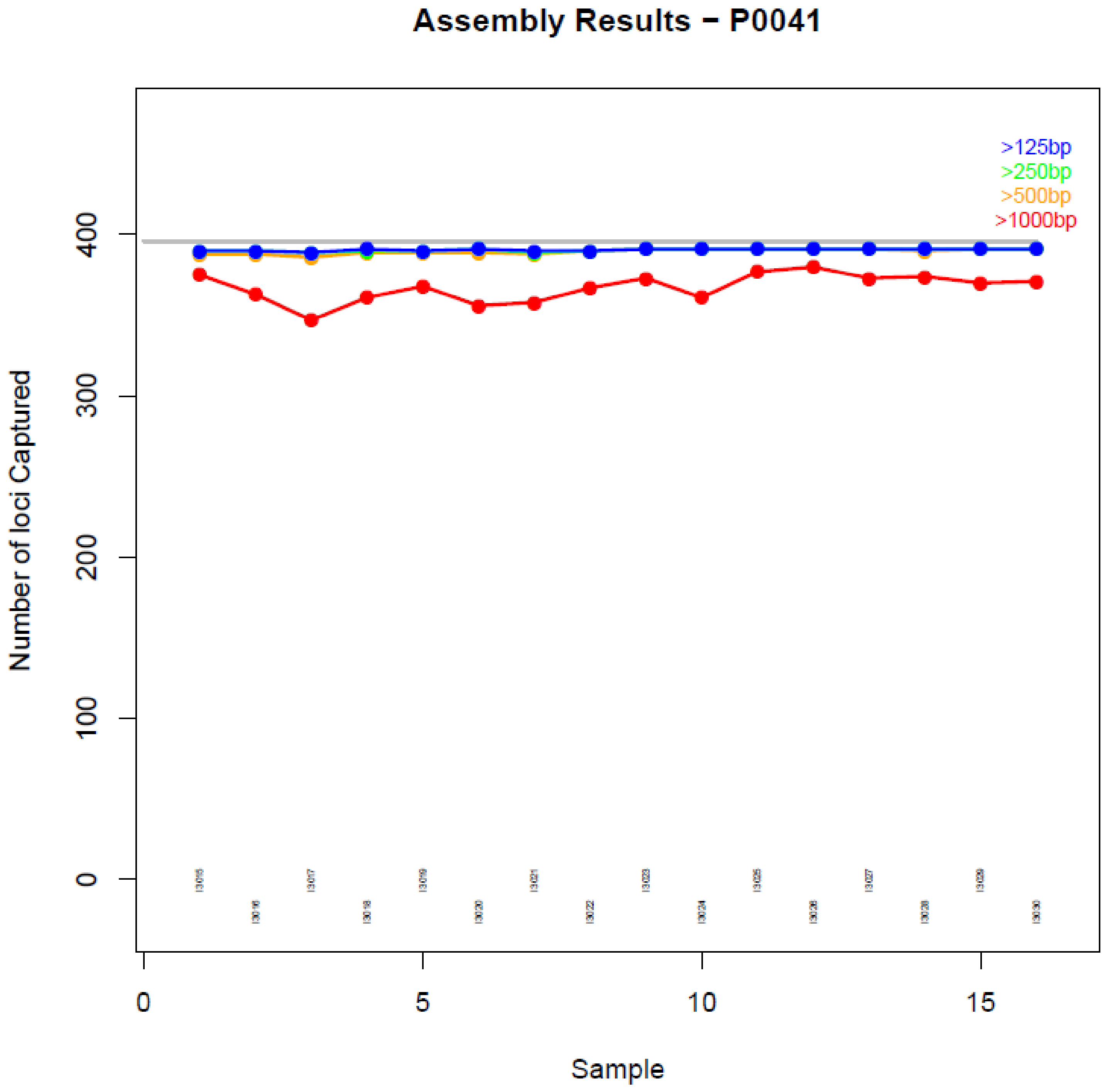
	H	I	J	K	L	M	N	C	P	Q	R	S	T	U	V	W	X	Y	Z
1	TaxonSE	nRawReads	nRawBases	nMergedRea	nMergedBase	nUnmergedR	nUnmergedB	E	nLoci125	nLoci250	nLoci500	nLoci1000	AvgLocLenAll	AvgLocLenUpper	%OnTarget	ReadsPerLocus	BasesPerLocus	AvgNHomologs	
2	I3015	4777138	896762700	1787929	388431857	1201280	360384000		390	390	388	375	1786,858561	1645,004963	24,3780062	2650,260546	452968,6749	1,49627792	
3	I3016	5862135	978836100	2599348	532361802	663439	199031700		390	390	388	363	1695,937965	1553,528536	24,840621	2449,873449	450825,5285	1,54094293	
4	I3017	1578221	264897900	695228	141742225	187765	56329500		389	389	386	347	1547,647643	1403,248139	30,3118736	798,7543424	148980,7717	1,25062035	
5	I3018	7158028	1157188800	3300732	662671015	556564	166969200		391	390	389	361	1650,307692	1511,124069	22,8521598	2518,937965	470448,4069	1,49875931	
6	I3019	6832014	1175247000	2914524	618704165	1002966	300889800		390	390	389	368	1729,26799	1585,047146	22,8918046	2845,146402	522361,4243	1,52605459	
7	I3020	8530258	1360788900	3994295	675067081	541668	162500400		391	391	389	356	1632,858561	1492,560794	16,2922831	1908,617866	338607,6055	1,59057072	
8	I3021	5324704	853784400	2478756	486518632	367192	110157600		390	389	388	358	1617,186104	1479,449132	22,9678376	1811,129032	340058,6303	1,4292804	
9	I3022	7783916	1265887200	3564292	725790044	655332	196599600		390	390	390	367	1707,275434	1558,615385	25,2503048	3128,148883	577931,0099	1,5955335	
10	I3023	5629711	963713100	2417334	508535533	795043	238512900		391	391	391	373	1777,593052	1629,409429	25,6548163	2629,223325	475567,9975	1,53101737	
11	I3024	7096102	1138941300	3299631	650125192	496840	149052000		391	391	391	361	1655,640199	1515,521092	24,7735938	2684,885856	491277,6948	1,62779156	
12	I3025	7002176	1224966300	2918955	603449691	1164266	349279800		391	391	391	377	1819,027295	1669,471464	22,1433948	2943,471464	523490,4541	1,60545906	
13	I3026	7078838	1323175500	2668253	600389055	1742332	522699600		391	391	391	380	1860,635236	1703,942928	21,6891425	3486,764268	604437,4665	1,60297767	
14	I3027	5776376	998519400	2447978	529884733	880420	264126000		391	391	391	373	1759,531017	1618,62531	30,9045738	3332,885856	608897,3524	1,57816377	
15	I3028	7048298	1290672000	2746058	609684588	1556182	466854600		391	391	390	374	1824,014888	1676,92804	22,2142558	3392,459057	593412,33	1,62034739	
16	I3029	4548558	785783400	1929280	418785734	689998	206999400		391	391	391	370	1734,349876	1590,493797	36,683912	3082,540943	569633,9156	1,55334988	
17	I3030	8052928	1328104500	3625913	752736503	801102	240330600		391	391	391	371	1750,191067	1604,126551	26,5514112	3546,009926	654276,2531	1,63027295	
18	I5712	5683720	986781900	2394447	509936184	894826	268447800		391	391	391	374	1785,699752	1641,181141	28,3027064	3224,774194	546659,3896	1,58064516	
19	I5713	6293010	1144346400	2478522	530904291	1335966	400789800		390	390	387	373	1870,684864	1722,054591	22,3012074	3281,292804	515580,7246	1,63523573	
20	I5714	4630225	755028900	2113462	427029470	403301	120990300		390	390	387	358	1625,885856	1495,853598	33,273895	2447,878412	452475,2432	1,52605459	
21	I5715	4364886	696579600	2042954	415050815	278978	83693400		391	391	388	356	1608,459057	1482,42928	45,3190482	3019,143921	560858,8859	1,57320099	
22	I5716	8352118	1440378000	3550858	745960878	1250402	375120600		390	390	389	370	1773,158809	1631,315136	20,4688158	3395,761787	569409,6849	1,72208437	
23	I5717	6539544	1103551800	2861038	586401280	817468	245240400		389	389	387	368	1732,682382	1590,8933	24,6082408	2896,967742	507822,3002	1,61042184	
24	I5718	7713896	1391672100	3074989	655068983	1563918	469175400		390	389	388	375	1884,523573	1737,153846	17,1303093	3013,734491	477882,2333	1,66253102	
25	I5719	3392535	622851300	1316364	272991928	759807	227942100		391	391	390	369	1798,952854	1663,37469	33,1683507	2575,054591	412286,737	1,54342432	
26																			
27																			
28																			
29																			

Description of the assembled data (example includes nine samples)

nGenes=434
Number of taxa = 9
Number of sites = 711063
Number of variable sites = 16986
Number of informative sites = 4373
Number of characters (total) = 6399567
% Missing characters (N's and -'s considered only) =1.4503637511725402
% Missing characters (all considered) =1.5992800762926618


Assembly Results − P0041
Consensus Sequence Length
4000
3000
2000
1000
0
I3015
I3017
I3019
I3021
I3023
I3025
I3027
I3029
0
5
10
15
Sample


Assembly Results − P0041
Number of Consensus Sequences
0
100
200
300
400
500
600
0
1000
2000
3000
4000
Length


Assembly Results − P0041
Number of loci Captured
0
100
200
300
400
>125bp
>250bp
>500bp
>1000bp
I3015
I3016
I3017
I3018
I3019
I3020
I3021
I3022
I3023
I3024
I3025
I3026
I3027
I3028
I3029
I3030
0
5
10
15
Sample

Approaches to analyze phylogenomic data

Important issues to consider:

- **!!!** amount of missing/ambiguous data
- **!!!** alignment
- **!** heterozygous data (phased vs. unphased!)

Concatenation versus species tree inference (coalescent analyses)

Concatenation:

- NJ – Geneious
- MP – TNT
- ML – RAxML (…), ~ FastTree
- BI – MrBayes, BEAST …

Coalescent-based Methods for Species Tree Inference

- Summary statistic methods: Start with estimated gene trees
  - Using estimated branch lengths:
    - STEM (Kubatko et al. 2009)
    - STAR, STEAC (Liu et al. 2009)
  - Using topology information only:
    - Minimize Deep Coalescences (PhyloNet; Than & Nakhleh 2009)
    - MP-EST (Liu et al. 2010)
    - ST-ABC (Fan and Kubatko 2011)
    - STELLS (Wu 2011)
- Methods that utilize the full data: Input is aligned sequences
  - BEST (Liu and Pearl 2007)
  - *BEAST (Heled and Drummond 2010)
  - New method based on algebraic statistics (Chifman and Kubatko 2013)

- Comparison of approaches:
  - Summary statistic methods
    - Advantage: Quick
    - Disadvantage: Ignore information in data
    - Most current implementations do not easily allow for assessment of uncertainty
  - Full data methods
    - Advantage: Fully model-based framework
    - Disadvantage: Computationally intensive, sometimes prohibitively so
    - Both BEST and *BEAST utilize a Bayesian framework and involve MCMC